

Machine Learning Approaches to Ensure Statistical Reproducibility of ESM Simulations

Salil Mahajan, Joe Kennedy, Kate Evans, Min Xu, Matt Norman, Michael Kelleher, Marcia Branstetter, Peter Caldwell, Andy Salinger

ORNL, LLNL, SNL

ORNL is managed by UT-Battelle, LLC for the US Department of Energy



Motivation:

- E3SM: Software and Algorithms (PI: Andy Salinger, SNL):
 - Effectively exploit DOE's leadership class HPC capabilities, improving model trust-worthiness
- Code Evolution:
 - Bit-for-bit reproducing changes
 - E.g. Adding a new compset, new output variable
 - Non-b4b changes
 - Different climate (statistics) expected
 - E.g. New parameterizations modules, new tunings
 - Same climate (statistics) expected
 - E.g. code porting, refactoring, GPU kernel, etc.
- Goal: Test the null hypothesis that climate simulation is similar for unintended nonb4b changes.



Motivation

- Truncated Floating Point arithmetic:
 - Round-off differences
 - Non-associative:
 - $(-1 + 1) + 2^{-53} \neq -1 + (1 + 2^{-53})$
 - Optimizations, hybrid architectures
- Climate models:

CAK RIDGE

3

- Chaotic, non-linear system
- Round-off differences grow quickly
- Problem: identify systematic bugs in non-BFB reproducible environment.



Lorenz attractor (Source:en.wikipedia.org/wiki/Chaos_theory)



WEATHER PATTERN WEDNESDAY 12/23

COOLER

E3SM Testing

E3SM Testing Suite (bfb): ٠

	41 passed, 0 failed, 0 not run, 0 missing.		
* APT (auto promotion test (default length)) * CME (compare met and comf interfaces (10 days))	Name A	Status 🛧	Time Details
* FRR (branch/exact restart test)	ERIO.ne30_g16_rx1.A.cori-knl_intel	Passed	24m 37s Completed (PASS
* EPU (hybrid/oxort rotati tot)	ERP_Ln9.ne4_ne4.FC5AV1C-L.cori-knl_intel	Passed	3m 33s Completed (PASS
* EPU (hybrid/branch/avact restart test, default 3+19/10+9/5+4 days)	ERS.f09_g16.I1850CLM45CN.cori-knl_intel	Passed	14m 4s Completed (PASS
* EPG (over the test from the test, default 6 dove + 5 dove)	ERS.f09_g16.I1850CLM45CN.cori-knl_intel.clm-bgcinterface	Passed	14m 31s Completed (PASS
* EPG (exact restart from startup, default 0 days + 5 days)	ERS.f09_g16.ICLM45BC.cori-knl_intel	Passed	10m 22s Completed (PASS
lafe deve = 1)	ERS.f09_g16_g.MALISIA.cori-knl_intel	Passed	5m 10s Completed (PASS
Info dbug = 1))	ERS.f19_f19.I1850CLM45CN.cori-knl_intel	Passed	5m 48s Completed (PASS
TCP (cice penormance test)	ERS.f19_f19.i20TRCLM45CN.cori-knl_intel	Passed	9m 52s Completed (PASS
LAR (long term archive test)	ERS.f19_f19.ICLM45.cori-knl_intel	Passed	5m 49s Completed (PASS
NCK (multi-instance validation vs single instance (default length))	ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-betr	Passed	6m 37s Completed (PASS
* NOC (multi-instance validation for single instance ocean (default length))	ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-vst	Passed	6m 33s Completed (PASS
* OCP (pop performance test)	ERS.f19_g16.I1850CNECACNTBC.cori-knl_intel.cim-eca	Passed	6m 1s Completed (PASS
* P4A (production branch test b40.1850.track1.1deg.006 year 301)	ERS.f19_g16.I1850CNECACTCBC.cori-knl_intel.cim-eca	Passed	6m 2s Completed (PASS
* PEA (single pe bfb test (default length))	ERS.f19_g16.I1850CNRDCTCBC.cori-knl_intel.clm-rd	Passed	5m 32s Completed (PASS
* PEM (pes counts mpi bfb test (seq tests; default length)) * PET (openmp bfb test (seq tests; default length))	ERS.119_g16.11850GSWCNPECACNTBC.cori-knl_intel.cim- eca_119_g16_11850GSWCNPECACNTBC	Passed	10m 9s Completed (PASS
* PFS (performance test setup)	ERS.119_g16.11850GSWCNPRDCTCBC.cori-knl_intel.cim- ctc_f19_g16_11850GSWCNPRDCTCBC	Passed	10m 6s Completed (PASS
default 6 days + 5 days)	ERS.119_g16.120TRGSWCNPECACNTBC.cori-knl_intel.clm- eca_119_g16_120TRGSWCNPECACNTBC	Passed	4m 56s Completed (PASS
* SBN (smoke build-namelist test (just run preview_namelist and check input data))	ERS.119_g16.120TRGSWCNPRDCTCBC.cori-knl_intel.clm- ctc_f19_g16_120TRGSWCNPRDCTCBC	Passed	4m 35s Completed (PASS
* SEC (acquiencing bits test (10 day acquience tests))	ERS.f19_g16_rx1.A.cori-knl_intel	Passed	5m 29s Completed (PASS
* SEQ (sequencing bibles) (10 day seq,conclests))	ERS.ne11_oQU240.I20TRCLM45.cori-knl_intel	Passed	7m 11s Completed (PASS
SWIS (shinke statup test (default length))	ERS.ne30_g16_rx1.A.cori-knl_intel	Passed	5m 22s Completed (PASS
Shore PCC)	ERS_Ld20.f45_f45.ICLM45ED.cori-knl_intel.clm-fates	Passed	35m 45s Completed (PASS
	ERS_Ld5.T62_oQU120.CMIPASO-NYF.cori-knl_intel	Passed	10m 24s Completed (PASS
	ERS_Ln5.ne4pg2_ne4pg2.FC5AV1C+L.cori-knl_intel.cam-thetahy_sl_pg2	Passed	5m 12s Completed (PASS

Non bit for bit changes: ٠

- Convergence test, perturbation growth test and climate reproducibility tests _
- Expert opinion, ad-hoc tests _

ii my.cdash.or E3SM CURRENT Dashboard Up Projec

Show Filters

Testing started on 1980-01-01 00:00:00

Site Name:cori-kni
Build Name:e3sm_developer_next_intel
Total time:5h 49m 25s
OS Name:Linux
OS Version:Commit: ca197b090117a93db4c5237bd836a1efa12a3a32Total testing time: 10443 seconds
Compiler Name:unknown
Compiler Version:unknown

Name 🔨	Status 🛧	Time	Details	History	o annary
ERIO.ne30_g16_rx1.A.cori-knl_intel	Passed	24m 37s	Completed (PASS)	Stable	Stable
ERP_Ln9.ne4_ne4.FC5AV1C-L.cori-knl_intel	Passed	3m 33s	Completed (PASS)	Stable	Stable
ERS.109_g16.11850CLM45CN.cori-knl_intel	Passed	14m 4s	Completed (PASS)	Stable	Stable
ERS.f09_g16.I1850CLM45CN.cori-knl_intel.clm-bgcinterface	Passed	14m 31s	Completed (PASS)	Stable	Stable
ERS.109_g16.ICLM45BC.cori-knl_intel	Passed	10m 22s	Completed (PASS)	Stable	Stable
ERS.f09_g16_g.MALISIA.cori-knl_intel	Passed	5m 10s	Completed (PASS)	Stable	Stable
ERS.f19_f19.I1850CLM45CN.cori-knl_intel	Passed	5m 48s	Completed (PASS)	Stable	Stable
ERS.f19_f19.i20TRCLM45CN.cori-knl_intel	Passed	9m 52s	Completed (PASS)	Stable	Stable
ERS.f19_f19.ICLM45.cori-knl_intel	Passed	5m 49s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-betr	Passed	6m 37s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CLM45.cori-knl_intel.clm-vst	Passed	6m 33s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CNECACNTBC.cori-knl_intel.clm-eca	Passed	6m 1s	Completed (PASS)	Stable	Stable
ERS.f19_g16.l1850CNECACTCBC.cori-knl_intel.clm-eca	Passed	6m 2s	Completed (PASS)	Stable	Stable
ERS.f19_g16.I1850CNRDCTCBC.cori-knl_intel.clm-rd	Passed	5m 32s	Completed (PASS)	Stable	Stable
ERS.f19_g16.11850GSWCNPECACNTBC.cori-knl_intel.clm- eca_f19_g16_11850GSWCNPECACNTBC	Passed	10m 9s	Completed (PASS)	Stable	Stable
ERS.119_g16.11850GSWCNPRDCTCBC.cori-knl_intel.clm- ctc_f19_g16_11850GSWCNPRDCTCBC	Passed	10m 6s	Completed (PASS)	Stable	Stable
ERS.f19_g16.l20TRGSWCNPECACNTBC.cori-knl_intel.cim- eca_f19_g16_l20TRGSWCNPECACNTBC	Passed	4m 56s	Completed (PASS)	Stable	Stable
ERS.119_g16.I20TRGSWCNPRDCTCBC.cori-knl_intel.cim- ctc_119_g16_I20TRGSWCNPRDCTCBC	Passed	4m 35s	Completed (PASS)	Stable	Stable
ERS.f19_g16_rx1.A.cori-knl_intel	Passed	5m 29s	Completed (PASS)	Stable	Stable
ERS.ne11_oQU240.I20TRCLM45.cori-knl_intel	Passed	7m 11s	Completed (PASS)	Stable	Stable
ERS.ne30_g16_rx1.A.cori-knl_intel	Passed	5m 22s	Completed (PASS)	Stable	Stable
ERS_Ld20.f45_f45.ICLM45ED.cori-knl_intel.clm-fates	Passed	35m 45s	Completed (PASS)	Stable	Stable
ERS_Ld5.T62_oQU120.CMPASO-NYF.cori-knl_intel	Passed	10m 24s	Completed (PASS)	Stable	Stable
ERS_Ln5.ne4pg2_ne4pg2.FC5AV1C+L.cori-knl_intel.cam-thetahy_sl_pg2	Passed	5m 12s	Completed (PASS)	Stable	Stable
NCK.f19_g16_nx1.A.cori-knl_intel	Passed	5m 37s	Completed (PASS)	Stable	Stable

CAK RIDGE

National Laboratory The main thing that distinguishes legacy code from non-legacy code is tests, or rather a lack of tests. Michael Feathersedit

Short Independent Simulation Ensemble

 $T'_j = (1+x')T_j$

x' is uniform random number transformed to range from (-10⁻¹⁴, 10⁻¹⁴)



National Laboratory

Short Independent Simulation Ensembles



Climate Reproducibility Tests: Ensemble Based Multivariate ML Approach

Accelerate and add rigor to the verification of E3SM for non-BFB changes

• Approach:

- Ensemble vs. ensemble
- Short (1yr) ensembles
- Short Ensembles:
 - Quantify natural variability
 - Computationally efficient (Mahajan et al. 2017)



- Leverage two sample equality of distribution tests from the ML community:
 - e.g. cross-match test, energy test, kernel test
 - Distribution-free/non-parametric
 - Effective at high dimensions, low sample sizes
 - Used widely in other fields, e.g. genetics, image processing, etc.

Short Independent Simulation Ensembles

- Packing simulations together is economical as compared to a SLR
- Compare a 100 1-yr ensemble vs. a 100-yr long run
 - Poor Weak and Strong Scaling for 100-yr long run smaller work load and increased MPI communications with increasing core counts
 - 100x greater workload per node for 100 member 1-yr ensemble on the same no. of nodes
 - Significantly reduced relative MPI and PCI-e overheads for ensembles:
 - Better parallel scaling
 - Faster throughput for ensembles:
 - Large core counts
 - Higher priority (capability scale) on leadership class machines (e.g. OLCF, NERSC, etc.)
 - Example (atmosphere spectral element 2 degree model):
 - Long run (100 years): 1536 elements, 96 nodes, 16 elements per node
 - SISE (100 1yr runs): 48 nodes each, 32 elements per node (total nodes: 4800)
- Usage:

CAK RIDGE

National Laboratory

- Solution reproducibility tests
- Scientific Applications



Courtesy: David Hall (https://www.earthsystemcog.org/projects/dcmip-2016/HOMME-NH)



Short Ensembles: Scientific Utility



Equality of Distribution Tests

- Energy Test (e.g. Szekely and Rizzo, 2004):
 - e-distance metric

$$e = \frac{nm}{n+m} \left(\frac{2}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \|X_i - Y_k\| - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|X_i - X_j\| - \frac{1}{m^2} \sum_{l=1}^{m} \sum_{k=1}^{m} \|Y_l - Y_k\| \right)$$

where X_1, \ldots, X_n and Y_1, \ldots, Y_m are the multivariate vectors of the baseline and perturbed ensembles.

- Small values of e indicate same population
- Derive null distribution by resampling





Equality of Distribution Tests

- Kernel Test (e.g. Gretton et al. 2006):
 - Maximum mean discrepancy (MMD) metric

$$MMD = \left(\frac{1}{n^2}\sum_{i,j=1}^n k(X_i, X_j) - \frac{2}{nm}\sum_{i,j=1}^{n,m} k(X_i, Y_j) + \frac{1}{m^2}\sum_{i,j=1}^m k(Y_i, Y_j)\right)^{\frac{1}{2}}$$

where k represents the kernel in its class of functions that maximizes MMD

- Small values of MMD indicates same population
- Derive null distribution by resampling



Equality of Distribution Tests

- Kolmogorov Smirnov (KS) Testing Framework:
 - Null Hypothesis (*H*₀): Two ensembles represent the same climate state.
 - Use global annual means of standard model output variables (158 variables).
 - H_0 : A variable between the two ensembles belong to the same distribution.
 - Test H_0 for each variable using a KS test.
 - Test statistic (t): No. of variables that reject H_0 at a given confidence level (say 95%).



- Test statistic (*t*): No. of variables that reject H_0 at a given confidence level (say 95%).
- *H*₀ rejected if *t* > *a*, where *a* is some critical number for a significance level (Type I error rate).
- a is empirically from an approximate null distribution of t derived using resampling techniques.

Significance Level (Type I Error Rate): Resampling

- Simulations from the two ensembles of size *n* and *m* are pooled together.
- Simulations from the pool are then randomly assigned to one of two groups of sizes *n* and *m*.

- The *t-statistic* is then computed for the random drawing.
- Repeat

CAK RIDGE

- If all possible random drawings are made, the null distribution of *t* is exact.
 - We conduct 500 drawings approximate null distribution.



Model Verification Using Ensembles: Known Climate Changing Perturbation

- Model: DOE E3SM v1
- Configuration: Active atmosphere land, prescribed cyclical F2000 SSTs and sea-ice distribution (FC5)
- Spatial Resolution: ~500km at the equator (5 degrees), 30 vertical layers
- Machine Configuration: PGI compiler on Titan
- Ensembles: Machine-precision level random perturbations to the initial 3-D temperature field
 - 30 member SISE
 - $T'_{j} = (1+x')T_{j}$, x' is random number transformed to range from (-10⁻¹⁴, 10⁻¹⁴)
- Turn a tuning parameter knob: zm_c0_ocn (control case: 0.007, modified: 0.045)



KS Testing Framework Results

Name	Description	Ens. Size
Default c0_ocn	Default model settings	30
Perturbed c0_ocn	Perturbed model parameter	30



	Comparison	Test Statistic (t)	Critical No.	H0 Test
¥	Default vs. perturbed c0_ocn	119	13	Reject

Power Analysis (Type II Error rate)

Type II error rate: Probability of accepting a false null hypothesis

- Turn a tuning parameter knob incrementally: zm_c0_ocn (0.007 to 0.045)
- Ensembles:
 - 100 members for each case
 - $T'_j = (1+x')T_j$, x' is random number transformed to range from (-10⁻¹⁴, 10⁻¹⁴)
- Power Analysis:
 - Randomly pick N=30 (=40, 50, 60) members from the control and perturbed sets
 - Conduct test
 - Repeat (500 times)



Power Analysis: KS Testing Framework

Controlled changes to zm_c0_ocn tuning parameter in Deep Convection



Example of Power Analysis.

Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case ($zm_co_ocn = 0.0070$) for different short simulation (1yr) ensemble sizes (N).

Power Analysis

Controlled changes to zm_c0_ocn (= 0.0070, defau in Deep Convection



Energy Test



18

Kernel Test



KS Testing Framework



Open slide Mahajan erallizo19



0.8



Kernel Test

375 350 325 300 200

dcs

dcs

325 300 200 100

KS Testing Framework



Open slide mahajahetal. 2019

CAK RIDGE

Vational Laboratory

Power Analysis: Atmosphere tests

- Expand on Power Analysis:
 - More tuning parameters
 - ice_sed_ai
 - sol_factb_interstitial
 - sol_factic_interstitial
 - cldfrc_dp1
 - zm_conv_Ind
 - dcs
 - zm_conv_ocn
 - zm_conv_dmpdz
- KS testing framework most powerful:
 - detects changes of smaller magnitudes confidently
 - compared to Kernel and Energy test.



Example of Power Analysis. Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a EAM tuning parameter from a control case (dcs = 400) for different short simulation (1yr) ensemble sizes (N).



Test Case: Cori vs. Edison

Evaluate if E3SMv1 DECK simulations on Edison can be reproduced on Cori

- Conducted short simulation (1yr) ensembles on both Edison and Cori:
 - F1850C5-CMIP6 compset
 - ne4 (100 ensemble members)
 - ne30 (30 ensemble members)
- All three TSC (Wan, et al.), perturbation growth (Singh, et al.), and KS climate reproducibility tests passed.
- Implications: Cori can be confidently used for remaining DECK simulations



News from DOE's state-of-the-science earth system model development project.



Can We Switch Computers?

Will the difference between simulated past and future climates be due to greenhouse gases or due to a change of DOE supercomputers? Thanks to a software modernization project, E3SM developers can answer this question and more. <u>Read more</u>.





Reproducibility Tests (EAM) on Master

First Previous 1 2 Next Last

- Nightly tests run on Cori (E3SM custom tests)
 - Time step convergence test
 - Perturbation growth test
 - KS testing framework
- On CDASH under E3SM_Customs_Tests
 - <u>https://my.cdash.org/index.php?project=E3SM</u>
 - All runs archived:
 - Large ne4 1yr F1850C5 ensemble available (>1000)

				iy.coust.org					
		01	RNL email Google	Facebook The Weat	her Channel				
ogin All Dashboards									
S/ ESSM			1.0001 0100	0.7					Dashboard Calendar
			Pher Con	241					
hours ago: 22 tests fail hours ago: 1 test failed hours ago: 2 tests faile hours ago: 50 warring hours ago: 6 tests faile e fuil feed	ed on ellem, integration, next, pgi on ellem, integration, master, intel d en ellem, ann, rohn, next, intel a introduced on scream, unit, twiste, full up, debug d on ellem, developer, next, gnu								
SM_Baselines 11 bui	las								[view
			Configure		Build		Test		
Site	Build Name	Error	Warn 🗸	Error	Warn	Not Run	Fall ¥	Pass	Start Time ¥
ompy	∆ e3sm_integration_next_pgi	0				0	22	64	2 hours ago
andiatoss3	A eSsm_integration_next_intel 9.4	0				0		n_{\star}	7 hours ago
парру	0 e3sm_developer_next_gnu > -	0	2.+2			0	•	37	6 hours ago
andiatoss3	D eSsm_integration_master_intel > ~	0	0			0	1,	63	4 hours ago
oor-kni	D edsm.prod_next_intel	0				0	0	3	9 hours ago
	A size boround maint 1 intel	0	141	U	0	0	0	4	11 hours ago
ion py		0	0			0	0	41	2 hours ago
and along 2	A boonta interaction part intel	0	0			0	0		7 hours ago
sandiaross3	A territor internation master intel	0	0			0	0	2	Shows ago
		0	0						0110018-020
ISM_Machine_Covera	ga ð buldu								(view
		Confi	guro	Build		Test			
Site	Build Name	Error	Warn	Error ¥	Warn	Not Run 🌱 🛛 Fail 🜱	Pass		Start Time 💙
nvi -	A edsm_integration_next_intel 9%	0	0	2	0	0 1	87		16 hours ago
ebop	∆ e3sm_integration_next_intel 9 ≈	D	0			4 5]	79		8 hours ago
heta	A e3sm_integration_next_intel 9 4	0	0			0 1	82		Jun 27, 2020 - 19:29 UTC
heta	A s3sm_hi_res_master_intel	D	0			0	2		Jul 08, 2020 - 22:27 UTC
or-haswell	A e3sm_developer_next_intel	0	0			0 0	43		1 hour ago
oori-kni	A edam_hi_res_master_intel	0	0			0 0,	3.		4 hours ago
sompy	0 e3sm_extra_coverage_master_intel	0	0			0 0	10		12 hours ago
00-101	Bissin_developer_next_inter	0	0			0 0	41		30/01,2020 - 19:01 UTC
3SM_Custom_Tests /	build								(view 1
35M_Custom_Tests 1	build		Configur		Nations	Test	Buc		(view 1
SSM_Custom_Tests 7	build Build Name		Configur Error	Warn 🗸	Not Pun	Test Fail ♥	Pass 1		(view 1 Start Time ❤
ISM_Custom_Tests / Site ori-kni	build Duild Name Duild Name		Configur Error 0	Warn 🗸	Not Pun 0	Test Fal ♥ 2	Poss 1		(view 1 Start Time ♥ 4 hours ago
Site Site or-kni ISM_SCREAM 19 Juni	Build Name Decide Name de		Configur Error 0	Warn 🌱	Not Pun 0	Text Fei ¥ 2	Pass 1	-	(view 1 Start Time ❤ 4 hours ago (view 1
Site Site Sof-Kni SSM_SCREAM 19 Journ	build Dadd Name S elow, stm, chilp, nect, ring 9 = Bridd Name	Con	Configure	Warn V Build	Not Bun 0	Test Pal ¥ 2 Test Not Res	Pass 1		(view 1 Start Time ♥ 4 hours ago (view 1 Start Time ♥
SIM_Custom_Tests 1 Site por-kmi SIM_SCREAM 19 Jouh Site	buld Debrin, sam, ridib, rient, ring (P=) de Debrin anseen motion range (P)	Con Error	Configure 0 Ngune 0	Warn Y 1 Build Error	Not Run 0 Warn V	Test Fai ♥ 2 Not Rus Fai ♥	Pass 1 Pass 10		(view 1 Start Time ❤ 4 hours ago (view 1 Start Time ❤ Jart 66 5000 - thest UTC
Site Site corr.knl SSM_SCREAM 19 Junit Site Site	but Budd Nore 5 alon, an, ofter, nor, yna 9 - 10 Budd Nore 5 alon, ynan, mare yn 9 - 6 alon, graen, mare yn 9 -	Con Error 0	Configur Error 0 Figure Warn ¥ 0	Warn ¥ 1 Build Error	Not Run 0 Warn ¥	Test Fall ♥ 2 Not Rust 0 1 0 0	Pass 1 Pass 10 24		(view 1 Start Time ♥ 4 hours ago (view 1 Start Time ♥ Jul 06, 2020 - 10:01 UTC 11 hours ago
ISM_Custom_Tests 7 Site or-kni ISM_SCREAM 79 Jouh Site or-kni chite	bult Buld News Setter an other sectors 9 = Setter somer, manzer, mill 9 = Setter somer, mill 9 = S	Con Error 0 0	Configur Error 0 Rigure 0	Warn ¥ 1 Build Enter 0	Not Run 0 Warn V 20 00	Test Fail ♥ 2 Tost Not Rus Fail ♥ 0 0 0 0	Pass 1 10 24 24		(view 1 Start Time ❤ 4 hours ago (view 1 Start Time ❤ Jul 06, 2020 - 10:01 UTC 11 hours ago 11 hours ago
SM_Custom_Yests 7 Site or-inni SM_SCREAM 79 built SMe or-inni chite chite steen	Autor Sectors are reference are a P- an Sectors are reference and a sector are a sector and a se	Con Error 0 0 0 0	Configur 0 Ngure 0 0	Warn V 1 Build Error	Not Run 0 Warn ¥ 50 60 50	Test Fall ♥ 2 Not Run Fall ♥ 0 1 0 0 0 0 0 0	Pass 1 10 24 24 24		jven t Start Time ♥ 4 hours ago Jvien t Start Time ♥ Jul 06, 2000 - 1000 UTC 11 hours ago 11 hours ago 11 hours ago 11 hours ago
Site Content of Site Content o	built Build News 5 cline, and r, oth, race, race 9 = 0 Build News 5 cline, somer, race, race 9 = 0 Subsect, somer, race, r	Con Error 0 0 0 0	Configure	Warn V Build Error 0 0	Not Run 0 Warm • 0 0 0 0	Test Fell V 2 2 Test Not Run 0 1 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24		jvier 1 Start Time ♥ 4 hours ago Start Time ♥ Jul 06, 2020 - 10:01 UTC 11 hours ago Jun 04, 2020 - 01:32 UTC Jun 04, 2020 - 01:32 UTC
SM_Custom_Tests 7 Site of-ion SM_SCREAM 79 built SM_SCREAM 79 built SM STR STR STR STR STR STR STR STR STR STR	Autor Caleformative rest and Pro- Andrew and rest and Pro- Build Name Build Name Buil	Con Error 0 0 0 0 0 0	Configue Error 0 Mare V 0	Warn V 1 Build Enter 0 0 0	Not Run 0 20 60 0 0	Text Fall ¥ 2 2 Not Rus Fall ¥ 0 1 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24 24		jvier Start Time * 4 hours ago Julios, 2000 - 1000 UTO 11 hours ago 11 hours ago 11 hours ago 10 hours ago
SSM_Custorn_Tests / Site cori-ioni SSM_SCREAM /29 Juli SSM_SCREAM /20 JULI SSM_SCREAM	huld Teach	Con Error 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur Error Agure Warn V O	Warn V Build Error 0 0 0	Not Run 0 50 50 50 50 50 50 50 50 50 50 50	Test Fol V 2 2 Not Run Fail V 0 1 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24 24 24 24 24 24		jvier 1 Start Time ♥ 4 hours ago Start Time ♥ Jul 06, 2020 - 10:01 UTC 11 hours ago 11 hours ago 11 hours ago 11 hours ago 10.6, 2020 - 0:35 UTD Jun 54, 2020 - 0:45 UTD Jun 54, 2020 - 0:45 UTD Jun 55, 2020 - 0:55 UTD
SBM_Custorn_Tests / Site corr-ivn/ SBM_SCREAM /9 built SBM SBM SBM SBM SBM SBM SBM SBM SBM SBM	bult Buld Name A data, attra data, and a data attra dattra data attra data attra data at	Con Error 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configue Error 0 0 1 1 1 1 1 1 0 0 0	Warn > 1 1 0 0 0 0 0 0 0	Not Run 0 50 50 50 0 50 0 0 0 0 0 0 0 0 0 0 0	Test 2 2 2 Not Rus Fall ¥ 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24 24 24 99		jview Start Time ♥ 4 hours ago Jul 08, 2000 - 10:01 UTO 11 hours ago 11 hours ago 11 hours ago 10 not sign 00:00 - 01:84 UTD Jun 18, 2000 - 01:84 UTD Shown ago 11 hours ago
ISM_Custom_Tests / ISMs Or-Iv1 ISMs ISMs Or-Iv1 ISMS ISMS Or-Iv1 ISMS ISMS Or-Iv1 ISMS	Autor Control Marcol Anno 2014 A chana, anto yodo, yoo ang 2014 A chana, anto yodo, yoo ang 2014 A chana, yoo ang	Com Broot 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur Error 0 4gurs Wars V 0 1 1 1 1 1 1 1 0 0 0	Wan >	Not Run 0 20 00 00 00 00 00 00 00 00 00 00 00 0	Test Foll ♥ 2 Not Run Fall ♥ 0	Pass 1 10 24 24 24 24 24 24 99 99		juw Start Time ♥ 4 hours ago Uulos, 2000 - 1061 UTC 11 hours ago 11 hours ago 10 0, 2000 - 1058 UTC Im Ki, 2000 - 1058
Site Control 1 Site Control 19 Durits Site Co	للملة العصم	Error 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur Emot D D D D D D D D D D D D D D D D D D D	Wan ♥ Errer 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Pain 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Text Pail ♥ 2 Not Rus Fail ♥ 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24 24 24 99 69		jówe Start Time ↓ 4 hours ago Start Time ↓ 4 dows ago 10 hours ago 11 hours ago
Std_Custern, Tester 7 Ste Orivit Sta St	Autor Adams and velocities and the form Adams and velocities and the form Adams and velocities and the form Adams and the form of the form Adams and the form	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur Error 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Warn ▼ 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Pain 0 20 30 30 30 50 50 50 50 50 50 50 50 50 50 50 50 50	Not Not 0 10 10 0 10 10 0 10 10 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 10 24 24 24 24 24 24 24 24 24 24 24 24 24		joer Read There V 4 hours ago Were V More And Mark Constant 11 hours ago 11 hours ago
Star Custom Tests 1 Sta ochry Star Star Star Star Star Star Star Star	Autor Settors and updates are 10 = 0 Settors and updates are 10 = 0 Settors and updates are 10 = 0 Settors are 10 = 0 Set	600 60 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	Configur Configurs	Warn V Build Draw 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Run 0 50 50 50 50 50 50 50 50 50 50 50 50 5	Not Pail W 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Pass 1 19 24 24 24 24 24 24 99 99 99 99		John 1 Andrew C Andrew C Bart Thew Y Mark Rev C Bart Thew Y Mark Rev C Bart Thew Y Mark Rev C Mark Rev C
SSM_Custern, Tester 7 State on-Tart SSM_COLLENT 19 build SSM_COLLENT 19 build SSM_COLLENT 19 build SSM_CUSTERN SSM	Antif California and angle of an angle of angle	Con C	Configue From 0 1 1 1 1 1 1 1 1 1 1 1 1 1	Warn V Dual Emer 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	North Pun 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Pail V 0 10 10 10 0 10 0 0 0 0	Pass 1 1 24 24 24 24 24 24 24 00 00 00 00 10 10 10 10 10 10 10 10 10	Pass	Jeen 1 Sector Sector S
SSM_Custom_Tests 7 Sta contain SSM_Contain 70 cutom SSM 70 cm	لل للل للل	Corr 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8	Confue Email 0 Maper 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Warn V Ener 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not fluin 0 Wars 9 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ран Рай V Лог Лам 0 0 0 0 0 0 0 0 0 0 0 0 0	Pees 10 24 24 24 24 24 24 24 24 24 24 24 24 24	Pass 64	June 1 2 4 000 200 2 2 2 2 2 2 2 2 2 2 2 2 2
Custom, Tesse 17 Star	لل للل لللل للل للل	Con 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Water V Budy Draw 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Net fain 0 Were V 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Fail W 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	700 1 10 24 24 24 24 24 24 24 24 24 24 20 0 0 0	Pass 61,7,	jeer 1 300 - 100
Custom, Teste 1 Star	Auto	Core 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Wen V 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Nor Nor Wares V 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Run Far V 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	700 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Pass 64, 77, 27,	Jeen 1 State Then V 4 hone age 11 hone ag
SSM_Custom Fee SSM_CUSTOM SSM_CUSTOM SSM SSM_CUSTOM SSM SSM_CUSTOM	Autor	Core 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Warn V 1 Bott 0	Not have 0	Not Part W 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	740 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Pass 64, 77, 27, 79,	Jeer 1 Sector Sector S
Status Teste 17 State State State		Con 0 0 0 0 0 0 0 0 0 0 0 0 0	Configur 0 0 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Run 0 Notes V 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Fail V 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Page 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Pass 64 77 78 1	June 1 Standard That A 4 Your Say Control Control Control 10 Your Say 10 Your Say 10 Your Say 10 Your Say 10 Your Say 10 Your Say 11 Your Say 12 Your Say 13 Your Say 14 Your
Star Crustern, Tests / Star	Autor	Core 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 8 8	Configue Terre 0 1 1 1 1 1 1 1 1 1 1 1 1 1	BAE 9 1	Nor Nor S	Not Run V 0	Page 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	Pass 61, 77, 11, 63	Jerrer 1 1 1 1 1 1 1 1 1 1 1 1 1
State of the second sec	Autor	Con 0 0 0 0 0 0 0 0 0 0 0 0 0	Conjuga Inter 0 1 1 1 1 1 1 1 1 1 1 1 1 1	Base Base Borr 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not have a Not V Not V Not Not Not Not Not Not Not Not Not Not	Not Fail V 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Page 1 1 2 2 3 4 3 4 3 4 4 3 4 4 3 4 3 4 3 4 3 4	Pass 64, 77, 77, 11, 11, 10, 3,	Juer 1
State State		Con 0 0 0 0 0 0 0 0 0 0 0 0 0	Configue 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Par Not S 0 0 0 0 0 0 0 0 0 0 0 0 0	Not Fail V Not Rue Fail V 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Para 1 1 24 24 24 24 24 24 24 24 24 24 24 24 24	Pass 54 77 77 1 1 0 3 3 43	June 1 3 4 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1



EVV:

CAK RIDGE

National Laboratory

- Extended Verification and Validation for Earth System Models (EVV):
 - Python based toolkit:
 - Runs control and perturbed ensembles
 - Post-processes model output
 - Conducts tests
 - Publishes results and auxiliary plots, tables



MPAS-O Reproducibility tests: Ensembles

- Generate ensembles:
 - 1. Low Res NYF Ocean run:
 - 240 km resolution (7153 cells)
 - Run to quasi-equilibrium pick base initial condition
 - Perturb initial condition to machine order precision:
 - Add perturbations to 3D temperature field initial condition
 - Save perturbed initial condition files
 - Use create_clone to generate ensembles:
 - each run reading a different perturbed initial condition file
 - 2. Pertlim capability for MPAS-O (near future):
 - Replicate capability within EAM to MPAS-O
 - Automatically perturb initial conditions
 - Generate ensembles by tweaking a namelist parameter.
 - Replicate multi-instance capability within EAM to MPAS.





MPAS-O Reproducibility tests: Approach

Larger Null Hypothesis: Control and perturbed ensembles belong to the same population

- Generate control and perturbed ensembles at QU240 resolution
- Evaluate 5 prognostic variables (Baker et al. 2016)
 - SSH, T, U, V, Salinity
 - Annual average of year 2.
- Ocean variability is spatially very heterogenous (as compared to the atmosphere):
 - Evaluate at each grid point.
- Conduct fine-grained null hypothesis tests at each grid point:
 - Two sample KS test: Popular non-parametric test
 - Cucconi test: Better power, rank based non-parametric test.



Growth of Round-off differences in MPAS-O



25 **CAK RIDGE** National Laboratory

Cucconi Test

• Test Statistic:

$$\mathrm{CUC} = rac{U^2+V^2-2
ho UV}{2(1-
ho^2)}.$$
U V ho

U: based on squared sum of ranks of samples in Ensemble A in the two sample pool of Ensembles A and B

V: based on squared sum of contrary-ranks of samples in Ensemble_vA in the pool.

 ρ : Correlation coefficient between U and V

- Larger test-statistic indicates that Ensemble A and B come from different populations.
- See also
 Popular in other fields like hydrology, quality control, etc. (e.g. Mukherjee and Marozzi et al. 2014)



MPAS-O Reproducibility Tests: Approach

Correct for simultaneous multiple null hypothesis tests (M grid points)

False Discovery Rate (FDR) approach (Wilks et al. 2006, Ventura et al. 2004):

- For single test, null hypothesis is rejected if:
 - Test statistic p-value (p) is less than a critical value, α (say 0.05): $p \le \alpha$
 - For *M* tests, αM would be rejected for true null hypotheses just by chance
- For multiple tests, FDR constrains critical value (α_{FDR}) for local hypothesis tests (H_0):

$$lpha_{FDR} = \max_{j=1,2,...,M} \{p_j: p_j \le lpha(j/M)\}$$
 p_j are sorted p-values of M tests

- Global Null Hypothesis Test (G_0): Reject if $p_i \le \alpha_{FDR}$ at any grid point.
- Robust for correlated tests e.g. spatial correlations (Wilks et a. 2006, Renard et al. 2008).
- Used in testing field significance



FDR Approach: Illustration



$$\alpha_{FDR} = \max_{j=1,2,...,M} \{ p_j : p_j \le \alpha(j/M) \}$$

FIG. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with $q = \alpha = 20\%$. The ordered *p*-values, $p_{(i)}$, are plotted against i/n, i = 1, ..., n, and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, re-₿ OAK RIDG^{spectively.} National Laborato

28

Ventura et al. 2004

MPAS-O Reproducibility Tests

Evaluate False Positive Rate:

Bootstrap with Control Ensemble (150 ensemble members):

- Randomly draw two samples with N=M=30 members
- Conduct KS test and Cucconi test for alpha = 0.05
- Repeat 500 times at alpha = 0.05

KS test:

95th percentile of the no. of cells rejecting the local null hypothesis (FDR) = 0 95th percentile of the no. of cells rejecting the local null hypothesis = 426

Cucconi test:

95th percentile of the no. of cells rejecting the local null hypothesis = 15 95th percentile of the no. of cells rejecting the local null hypothesis = 643

MPAS-O Reproducibility Tests: Results

Known Climate Changing Case: GM Kappa = 600 (Default = 1800) 30 member ensembles for test and control case



Growth of machine precision differences in oQU240 MPAS-O and ensemble spread: L1 Norm (sum of absolute difference at each grid point, log-scale) of SST of each of the 100 ensemble members with round off differences in initial conditions compared to a reference run for the control (kappa = 1800, red lines) and modified (kappa = 600, blue lines) ensembles.

CAK RIDGE

30

Both tests reject the null hypothesis that the two ensembles belong to the same population at the 0.05 significance level.

MPAS-O Reproducibility Tests: Power Analysis

Type II error rate: Probability of accepting a false null hypothesis

- Turn a tuning parameter knob incrementally:
 - Gent and McWilliams kappa (600 to 1800):
- Ensembles:
 - 100 members for each case
 - $T'_j = (1+x')T_j$, x' is random number transformed to range from (-10⁻¹⁴, 10⁻¹⁴)
- Power Analysis:
 - Randomly pick N=30 (=40, 50, 60) members from the control and perturbed sets
 - Conduct test
 - Repeat (500 times)



MPAS-O Reproducibility Tests: Power Analysis

Controlled changes to GM kappa tuning parameter in MPAS-O



Power Analysis. Probability of correctly rejecting a false null hypothesis (Power) of the test in detecting changes to a MPAS-O tuning parameter from a control case (GM kappa = 1800) for different ensemble sizes (N).

Summary:

- Use short ensembles for model verification as E3SM adapts for Exascale
- Developed a multivariate testing framework for climate reproducibility after perturbation growth:
 EVV
- Power Analysis of tests to evaluate their detection limits
- Test Cases:
 - Known climate changing perturbations: tuning parameter changes
 - Compiler optimization choices, reproducibility of frozen model after months of software updates
 - Machine port from NERSC's Edison to Cori of E3SMv1 atmosphere model
- Expanding to include reproducibility testing to MPAS-O
 - Generated control and perturbed GMPAS-NYF ensembles using create_clone
 - KS Test and Cucconi tests with false discovery rates
 - Power Analysis with GM kappa tuning parameter



Next Steps and Challenges

- Future work for MPAS-O tests:
 - Conduct ensembles trajectories from a better quasi-equilibrium initial state
 - Power analysis with other controlled changes
 - Evaluate applicability of low-resolution results at high-resolution
 - Explore other multivariate tests
 - Apply to prior known non-b4b changes and live non-b4b changes
- Integrating tests into EVV/CIME.
- Develop ensemble-based tests for individual software kernels: RRTMGP, MG2, CLUBB, MAM4, etc. (in a SCM framework?)
- Investigate applicability to other model components.





Hack and Pedretti (2000)



Thanks!

• Acknowledgements:

- DOE E3SM Project and CMDV-SM Project
- Oak Ridge Leadership Computing Facility (OLCF)
- NERSC

(OAK RIDGE | cak ridge leadership National Lebatairy | computing facility



35



Test for Extremes

- Distribution tests perform poorly on distribution with different tails
 - Known for univariate tests, unexplored for multivariate tests.
- Use Generalized Extreme Value (GEV) theory (e.g. Mahajan et al. 2015, Evans et al. 2014).



Time (days)





where μ , σ and ξ represent the location, scale and shape parameter respectively.



Climate Extremes Test

- Null Hypothesis (G₀): Simulation of extremes of a variable between two SISE is statistically indistinguishable.
- Annual maxima for each grid point are fit to a GEV distribution.
- G_0 : Extremes at each grid point are statistically indistinguishable
- Test statistic (g): No. of grid points that reject G_0
- G_0 rejected if t > b, where b is some critical number, obtained using resampling techniques.



Climate Extremes



Climate Extremes

39

Comparison	Variable	Test statistic (g)	Critical value (β)	G ₀ Test
SISE-DEFAULT vs.	Precipitation Rate	5.1%	6.5%	Accept G_0
SISE-O1				
	Surface Temperature	5.0%	9.6%	Accept G_0
SISE-DEFAULT vs.	Precipitation Rate	4.7%	6.3%	Accept G_0
SISE-FAST				
	Surface Temperature	3.6%	9.6 %	Accept G_0
SISE-O1 vs. SISE-	Precipitation Rate	5.2%	6.5%	Accept G_0
FAST				
	Surface Temperature	10.3%	9.8%	Reject G ₀

- All SISE simulations are identical to each other in terms of their simulation of climate extremes.
- The result is in contrast to the result of the KS-testing framework.
- It suggests that either optimization choices do not effect climate extremes, or COAK RIDGE



• SLR is clearly distinct from the SISE-DEFAULT

KS Testing Framework Results

Comparison	Test Statistic (<i>t</i>)	Critical Value (α)	${ m H}_0$ Test Result
SLR vs. SISE-DEFAULT	80 (50.6 %)	15	Reject H ₀
SLR vs. SISE-LND-INIT	74 (48 %)	13	Reject H ₀



SLR vs. SISE

• Atmospheric models show that free atmospheric-only internal variability can include variability on longer time-scales (*e.g. James and James, 1989, Lorenz, 1990, Held, 1993, Marshall and Molteni, 1993*).



Atmospheric Low-frequency Variability

CAK RIDGE National Laboratory

42



Multivariate Cross-Match Test

- *n* 1-yr control runs (~C)
- *m* 1-yr modified runs (~M)
- Coarse grained: global annual means
- Multivariate vector for each run (size ~130)
- Pool vectors, N = n + m
- Pair vectors based on min.
 Mahalanobis distance
- $H_0: C = M$





Illustration of cross matching for a bivariate case with n = m = 10. *(Ruth, 2014)*

Cross-Match Test

• Null distribution of T-statistic:

$$P(T = a_1) = \frac{2^{a_1} (N/2)!}{\binom{N}{n} (\frac{n-a_1}{2})! a_1! (\frac{m-a_1}{2})!}$$



- i.e. when both samples belong to the same population
- where a_1 is the no. of pairs with one control and one perturbed vector
- Based on simple combinatorial arguments, thus exact
- Analogous to the probability of drawing one red and one green ball

Single Long Runs: Scalability

- To enhance throughput, use more cores:
 - 5 simulated years per day (required)
- But, scaling (weak or strong) is not perfect:
 - Less work per core with large core counts
 - Increase in MPI communications
 - Smaller MPI messages
 - Large MPI latency
- MPI cost: 90%



Courtesy: Mark Taylor, AMWG meeting





Climate State Approach

- Several years of a control run
 - scientifically validated on a trusted machine
- Several years of the perturbed run
- Expert opinion from a subjective evaluation of plots, tables, etc.
- Expensive, slow and subjective, no quantitative standardized metric or cost function analysis.
- Need for tests for the multivariate problem of climate model verification.

CAK RIDGE

46

National Laboratory



Test Case: Optimization Choices

- Model: DOE E3SM v0.4
- Configuration: F1850C5
- Spatial Resolution: 208km at the equator (2 degrees), 30 vertical layers
- Machine Configuration: PGI compiler on Titan

KS Testing Framework Results

Comparison	Test Statistic (t)	Critical Value (α)	H ₀ Test
SISE-DEFAULT vs. SISE-O1	1 (0.6%)	17	Accept H ₀
SISE-DEFAULT vs. SISE-FAST	24 (15.2%)	14	Reject H ₀
SISE-O1 vs. SISE-FAST	23 (14.6%)	16	Reject H ₀

Aggressive compiler choices (SISE-FAST) with the PGI compiler on Titan can result in climate-changing simulations.



Test Case: Model Verification Using Ensembles: Frozen model configuration v0 vs. v1

- Configuration: F1850C5 compset (frozen after v0 bug-fixes, v0.4)
- Spatial Resolution: 208km at the equator (2 degrees), 30 vertical layers
- Goal: Evaluate if efforts towards exascale computing impact climate reproducibility:
 - New scientific features, code refactoring
 - CIME (Common Infrastructure for Modeling the Earth System) update
 - Compiler and Software library updates

Name	Ens. Size	CIME	PGI	p-netcdf
v0.4-2015	30	4.0	15.3	1.5.0
master	30	5.0	17.5	1.7.0
v0.4	27	4.0	17.5	1.7.0



Frozen model configuration v0 vs. v1

Comparison	Test Statistic (t)	Critical no. (α)	H0 Test
v0.4-2015 vs. master	6 (3.6%)	13	Accept H0
v0.4 vs. master	8 (4.2%)	13	Accept H0
v0.4-2015 vs. v0.4	5 (3%)	13	Accept H0

Software infrastructure updates are not climate changing. Frozen model configuration reproducible!

